Basics of Parameter Estimation in Probabilistic Models

Piyush Rai
IIT Kanpur

Probabilistic Machine Learning (CS772A)

Jan 11, 2016
Parameter Estimation

- Given: data $X = \{x_1, x_2, \ldots, x_N\}$ generated i.i.d. from a probabilistic model

$$x_n \sim p(x|\theta) \quad \forall n = 1, \ldots, N$$

- Goal: estimate parameter $\theta$ from the observed data $D$

- First, recall the Bayes rule: The posterior probability $p(\theta|X)$ is

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int_\theta p(X|\theta)p(\theta)d\theta} = \frac{\text{likelihood } \times \text{ prior}}{\text{marginal probability}}$$

- $p(X|\theta)$: probability of data $X$ (or “likelihood”) for a specific $\theta$

- $p(\theta)$: prior distribution (our prior belief about $\theta$ without seeing any data)

- $p(X)$: marginal probability (or “evidence”) - likelihood averaged over all $\theta$’s (also normalizes the numerator to make $p(\theta|X)$ a probability distribution)
Perhaps the simplest (but widely used) parameter estimation method

Finds the parameter \( \theta \) that maximizes the likelihood \( p(X|\theta) \)

\[
\mathcal{L}(\theta) = p(X|\theta) = p(x_1, \ldots, x_N | \theta) = \prod_{n=1}^{N} p(x_n | \theta)
\]

Note: Likelihood is a function of \( \theta \)
Maximum Likelihood Estimation (MLE)

- MLE typically maximizes the log-likelihood instead of the likelihood (doesn’t affect the estimation because log is monotonic)

- Log-likelihood:
  \[
  \log L(\theta) = \log p(X | \theta) = \log \prod_{n=1}^{N} p(x_n | \theta) = \sum_{n=1}^{N} \log p(x_n | \theta)
  \]

- Maximum Likelihood parameter estimation
  \[
  \hat{\theta}_{MLE} = \arg \max_{\theta} \log L(\theta) = \arg \max_{\theta} \sum_{n=1}^{N} \log p(x_n | \theta)
  \]
MLE: Consistency

- If the assumed model $p(x|\theta)$ has the same form as the true underlying model, then the MLE is consistent as the number of observations $N \to \infty$

$$\hat{\theta}_{MLE} \to \theta_*$$

where $\theta_*$ is the parameter of the true underlying model $p(x|\theta_*)$ that generated the data

- A rough informal proof: In the limit $N \to \infty$

$$\mathcal{L}(\theta) = \mathbb{E}_{x \sim p(x|\theta_*)}[\log p(x|\theta)]$$

$$= -\text{KL}(p(x|\theta_*)||p(x|\theta)) + \mathbb{E}_{x \sim p(x|\theta_*)}[\log p(x|\theta_*)]$$

(proof on the board)

- Thus $\hat{\theta}_{MLE}$, the maximizer of $\mathcal{L}(\theta)$, minimizes the KL divergence between $p(x|\theta_*)$ and $p(x|\theta_*)$. Since both have the same form, $\theta = \theta_*$
MLE via a simple example

- Consider a sequence of $N$ coin tosses (call head = 0, tail = 1)
- Each outcome $x_n$ is a binary random variable $\in \{0, 1\}$
- Assume $\theta$ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(x_n \mid \theta)$ is Bernoulli: $p(x_n \mid \theta) = \theta^{x_n}(1 - \theta)^{1-x_n}$
- Log-likelihood: $\sum_{n=1}^{N} \log p(x_n \mid \theta) = \sum_{n=1}^{N} x_n \log \theta + (1 - x_n) \log(1 - \theta)$
- Taking derivative of the log-likelihood w.r.t. $\theta$, and setting it to zero gives
  \[
  \hat{\theta}_{MLE} = \frac{\sum_{n=1}^{N} x_n}{N}
  \]
- $\hat{\theta}_{MLE}$ in this example is simply the fraction of heads!
- MLE doesn’t have a way to express our prior belief about $\theta$. Can be problematic especially when the number of observations is very small (e.g., suppose we only observed heads in a small number of coin-tosses).
Maximum-a-Posteriori Estimation (MAP)

- Allows incorporating our prior belief (without having seen any data) about $\theta$ via a prior distribution $p(\theta)$
- $p(\theta)$ specifies what the parameter looks like \textit{a priori}
- Finds the parameter $\theta$ that maximizes the \textit{posterior probability} of $\theta$ (i.e., probability in the light of the observed data)

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\theta|X)$$
Maximum-a-Posteriori (MAP) Estimation

- Maximum-a-Posteriori parameter estimation: Find the $\theta$ that maximizes the (log of) posterior probability of $\theta$

$$\hat{\theta}_{MAP} = \arg \max_\theta p(\theta|X) = \arg \max_\theta \frac{p(X|\theta)p(\theta)}{p(X)}$$

$$= \arg \max_\theta p(X|\theta)p(\theta)$$

$$= \arg \max_\theta \log p(X|\theta)p(\theta)$$

$$= \arg \max_\theta \{\log p(X|\theta) + \log p(\theta)\}$$

$$\hat{\theta}_{MAP} = \arg \max_\theta \{\sum_{n=1}^{N} \log p(x_n|\theta) + \log p(\theta)\}$$

- Same as MLE except the extra log-prior-distribution term!

- Note: When $p(\theta)$ is a uniform prior, MAP reduces to MLE
Let’s again consider the coin-toss problem (estimating the bias of the coin).

Each likelihood term is Bernoulli: \( p(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n} \)

Since \( \theta \in (0, 1) \), we assume a Beta prior: \( \theta \sim \text{Beta}(\alpha, \beta) \)

\[
p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}
\]

\( \alpha, \beta \) are called hyperparameters of the prior.
MAP via a simple example

- The log posterior probability = \( \sum_{n=1}^{N} \log p(x_n|\theta) + \log p(\theta) \)

- Ignoring the constants w.r.t. \( \theta \), the log posterior probability:
  \[
  \sum_{n=1}^{N} \{ x_n \log \theta + (1 - x_n) \log(1 - \theta) \} + (\alpha - 1) \log \theta + (\beta - 1) \log(1 - \theta)
  \]

- Taking derivative w.r.t. \( \theta \) and setting to zero gives
  \[
  \theta_{MAP} = \frac{\sum_{n=1}^{N} x_n + \alpha - 1}{N + \alpha + \beta - 2}
  \]

- Note: For \( \alpha = 1, \beta = 1 \), i.e., \( p(\theta) = Beta(1, 1) \) (which is equivalent to a uniform prior), we get the same solution as \( \theta_{MLE} \)

- Note: Hyperparameters of the prior (in this case \( \alpha, \beta \)) can often be thought of as “pseudo-observations”. E.g., in the coin-toss example, \( \alpha - 1, \beta - 1 \) are the expected numbers of heads and tails, respectively, before seeing any data.
Point Estimation vs Full Posterior

- Note that MLE and MAP only provide us with a best “point estimate” of $\theta$
  - MLE gives $\theta$ that maximizes $p(X|\theta)$ (likelihood, or probability of data \textit{given} $\theta$)
  - MAP gives $\theta$ that maximizes $p(\theta|X)$ (posterior probability of the parameter $\theta$)

- MLE does not incorporate any prior knowledge about parameters
- MAP does incorporate prior knowledge but still only gives a point estimate

Point estimate doesn’t capture the uncertainty about the parameter $\theta$

The full posterior $p(\theta|X)$ gives a more complete picture (e.g., gives an estimate of uncertainty in the learned parameters, gives more robust predictions/uncertainty in predictions, and many other benefits that we will see later during the semester)
Point Estimation vs Full Posterior

- Estimating (or “inferring”) the full posterior can be hard in general.

- In some cases, however, we can analytically compute the full posterior (e.g., when the prior distribution is “conjugate” to the likelihood).

- In other cases, it can be approximated via approximate Bayesian inference (more on this later during the semester).
Let’s come back once more to the coin-toss example

Recall that each likelihood term was Bernoulli: \( p(x_n | \theta) = \theta^{x_n}(1 - \theta)^{1-x_n} \)

The prior \( p(\theta) \) was Beta: \( p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1} \)

The posterior is given by

\[
p(\theta | X) \propto \prod_{n=1}^{N} p(x_n | \theta) p(\theta)
\]

\[
\propto \theta^{\alpha+\sum_{n=1}^{N} x_n-1}(1 - \theta)^{\beta+N-\sum_{n=1}^{N} x_n-1}
\]

It can be verified (exercise) that the normalization constant in the above is a Beta function \( \frac{\Gamma(\alpha+\sum_{n=1}^{N} x_n)\Gamma(\beta+N-\sum_{n=1}^{N} x_n)}{\Gamma(\alpha+\beta+N)} \)

Thus the posterior \( p(\theta | X) = \text{Beta}(\alpha+\sum_{n=1}^{N} x_n, \beta+N-\sum_{n=1}^{N} x_n) \)

Here, the posterior has the same form as the prior (both Beta)

Also very easy to perform online inference (posterior can be used as a prior for the next batch of data)
Assume starting with a uniform prior (equivalent to Beta(1,1)) in the coin-toss example and observing a sequence of heads and tails.
Conjugate Priors

- If the prior distribution is conjugate to the likelihood, posterior inference is simplified significantly.

- When the prior is conjugate to the likelihood, posterior also belongs to the same family of distributions as the prior.

- Many pairs of distributions are conjugate to each other. E.g.,
  - Bernoulli (likelihood) + Beta (prior) $\Rightarrow$ Beta posterior
  - Binomial (likelihood) + Beta (prior) $\Rightarrow$ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) $\Rightarrow$ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) $\Rightarrow$ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) $\Rightarrow$ Gamma posterior
  - and many other such pairs.

- Easy to identify if two distributions are conjugate to each other: their functional forms are similar. E.g., multinomial and Dirichlet

  \[
  \text{multinomial} \propto p_1^{x_1} \cdots p_K^{x_K}, \quad \text{Dirichlet} \propto p_1^{\alpha_1} \cdots p_K^{\alpha_K}
  \]
Conjugate Priors and Exponential Family

- Recall the exponential family of distributions

  \[ p(x|\theta) = h(x)e^{\eta(\theta)^\top T(x) - A(\theta)} \]

- \( \theta \): parameter of the family. \( h(x), \eta(\theta), T(x), \) and \( A(\theta) \) are known functions
- \( p(.) \) depends on data \( x \) only through its \textbf{sufficient statistics} \( T(x) \)
- For each exp. family distribution \( p(x|\theta) \), there is a conjugate prior of the form

  \[ p(\theta) \propto e^{\eta(\theta)^\top \alpha - \gamma A(\theta)} \]

  where \( \alpha, \gamma \) are the hyperparameters of the prior
- Updated posterior: posterior will also have the same form as the prior

  \[ p(\theta|x) \propto p(x|\theta)p(\theta) \propto e^{\eta(\theta)^\top [T(x)+\alpha]-[\gamma+1]A(\theta)} \]

- Updates by adding the sufficient statistics \( T(x) \) to prior’s hyperparameters
Next Class:
Probabilistic Linear Regression